arXiv:2303.05322v1 [cs.SD] 9 Mar 2023

通过文本转语音数据增强提升说话人脸系统的小样本学习能力

*Qi Chen[1], Ziyang Ma[1], Tao Liu[1], Xu Tan[2], Qu Lu[3], * Xie Chen[1], * Kai Yu[1]*

[1]X-LANCE Lab，计算机科学与工程系，教育部人工智能重点实验室，人工智能研究院，上海交通大学，中国[2]微软亚洲研究院，[3]上海媒体科技{cq1073554383,zym.22,liutaw,chenxie95,kai.yu}@sjtu.edu.cn

摘要

近年来，音频驱动的说话人脸受到学术界和工业界的广泛关注。然而，音频驱动的说话人脸的数据采集和标注耗费大量人力且成本高昂。数据资源的匮乏导致合成效果不佳。为缓解这一问题，我们提出使用文本转语音（TTS）进行数据增强，以提升说话人脸系统的小样本能力。我们引入软动态时间规整，解决文本转语音音频带来的错位问题；据我们所知，这是该方法首次应用于说话人脸任务。此外，我们还探索了 HuBERT 提取的特征，以利用音频的潜在信息，发现其效果优于其他特征。与基线模型相比，所提出的方法在均方误差分数、动态时间规整分数和用户研究偏好方面分别取得了 17%、14% 和 38% 的优势，这表明通过文本转语音数据增强提升说话人脸系统小样本学习能力的有效性。

索引词— 小样本学习，软动态时间规整，HuBERT，低资源

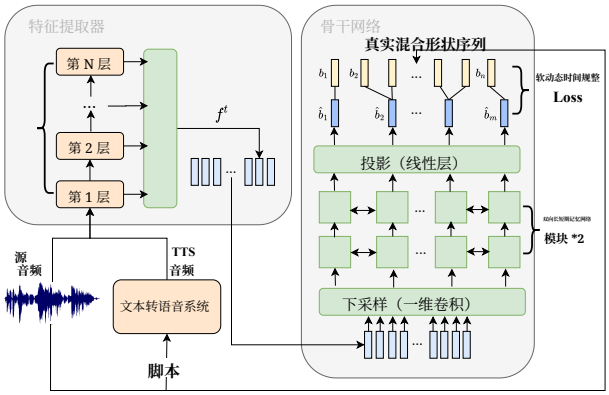


图 1。所提出的文本转语音增强的音频驱动说话人脸的整体架构。

1. 引言

近年来，数字人已成为学术界和工业界的热门话题。数字人有着众多应用，尤其是在媒体与娱乐行业，包括虚拟新闻主播、虚拟 YouTuber、虚拟偶像等。要生成高质量的数字人，合成自然的面部动作至关重要，也就是实现说话人脸生成。说话人脸主要分为两类：视频驱动和音频驱动。视频驱动的说话人脸也称为面部重演，重点是利用参考视频驱动肖像 [1, 2, 3]。然而，视频驱动的说话人脸有时会出现面部动作重演不稳定的问题。音频驱动的说话人脸旨在合成与给定音频相对应的面部动作，在面部动作合成方面更加稳定 [4, 5, 6]。

音频驱动的说话人脸中的面部动作表示是系统层面的一个关键因素。处理流程中常用的面部动作表示包括面部网格等基于关键点的格式 [7, 8] 或混合形状等基于运动的格式 [5]。获取这类语料库颇具挑战，因为它们是由专业设备录制的，这限制了其广泛应用。虽然有些移动设备能够方便地录制此类数据，但由于设备精度有限，采集到的数据并不准确。因此，在训练前需要对数据进行人工修整，这会耗费大量人力成本。如何通过小样本学习进行训练仍有待研究。

随着文本转语音（TTS）技术的发展， [9, 10, 11], 现成的 TTS 系统能够合成高质量音频，并常用于数据增强，以提升各种语音相关任务中的说话人多样性，例如语音识别 [12]和说话人验证 [13]。此外，基于大量无标注音频数据训练的预训练模型（PTMs）在音频特征提取方面展现出巨大潜力。因此，结合文本转语音数据增强与预训练模型，是提升低资源说话人脸生成效果的一种可行方法。

本文中，为缓解上述问题并提升说话人脸的鲁棒性，我们采用文本转语音增强策略生成额外的训练数据，并将其与少量有标注数据结合。为解决新生成数据中存在的错位问题，我们引入软动态时间规整 [14], ——动态时间规整（DTW） [15] 算法的一种可微形式——来对齐训练数据。据我们所知，这是该损失函数首次应用于说话人脸任务。随后，我们研究了利用 HuBERT [16] 模型提取鲁棒音频特征的方法。我们采用特征的加权和，以充分利用音频的底层信息。实验表明，在一个仅包含随机挑选的 10 条、时长约一分钟的语句的小型训练语料库上，我们的方法取得了显著提升。代码和演示现已发布，我们希望本方法能为相关领域作出贡献。

* Xie Chen 和 Kai Yu 是通讯作者。

† 代码和演示可在此处获取：https://github.com/Moon0316/T2A

2. 相关工作

2.1. 音频驱动的说话人脸

音频驱动的说话人脸旨在生成与输入语音信号相匹配的面部动作。文献中有两种音频驱动的说话人脸系统：一种是模块化系统，另一种是端到端系统。尽管端到端说话人脸系统方面已有积极的研究投入，并取得了令人鼓舞的进展 [17], 但迄今为止，模块化系统仍是实践中使用的主流系统，也是本文关注的重点。模块化系统的典型处理流程包含两个模块：首先将声学特征映射为某种面部运动表示，然后根据预测的面部动作表示渲染视频。以往的研究中，声学特征和面部动作表示有多种选择。Thies 等人 [4] 利用 DeepSpeech [18] 提取音频特征，随后采用基于卷积神经网络的模型，将音频特征映射为三维可变形模型（3DMM）特征，并以此生成神经纹理，最后采用图像到图像转换方法渲染出逼真的视频。Chen 等人 [5] 使用音素后验概率图（PPGs）、基频和能量作为音频特征，通过混合专家（MOE）Transformer 将其映射为混合形状，最后由渲染引擎生成动画。Zhou 等人 [6] 通过语音转换模型 AutoVC [19] 将音频中的语言信息与说话人信息解耦，并应用长短期记忆网络预测面部关键点，最后通过图像到图像转换合成逼真的视频，或通过人脸变形合成动画。

2.2. HuBERT

随着互联网上原始音频数据量的爆炸式增长，如何有效利用这些无标签数据已成为研究热点。采用自监督学习的预训练模型（PTM）在一系列语音相关任务中取得了巨大成功 [20, 21, 22, 16, 23, 24]。其中，HuBERT 是语音表征学习领域表现最佳的模型之一。在预训练阶段，HuBERT 通过采用类似 BERT 的预测损失来预测离散目标标签。离散标签是通过对原始音频本身进行 K-means 聚类得到的，无需使用任何有标注数据。在微调阶段，采用监督学习使模型适应特定的下游任务，例如自动语音识别（ASR）。近期的 SUPERB 基准测试 [25] 表明，在各种下游任务中，从自监督预训练模型提取的特征普遍优于传统的手工声学特征，例如梅尔频率倒谱系数。

2.3. 动态时间规整

动态时间规整（DTW） [15] 是一种动态规划算法，用于确定两组序列数据的最佳对齐方式并量化其相似程度。DTW 可通过扭曲时间序列的时间轴并执行动态规划算法，计算长度不同的两个序列之间的最小距离。它最初旨在解决口语词识别任务中相同说话模式下的语速变化问题 [15], ，并成功量化了长度可变的语音序列与词模板之间的相似度。Donald 等人 [26] 将 DTW 用于模式匹配问题，并发现 DTW 在从时间序列中寻找模式方面展现出巨大潜力。

3. 方法

3.1. 文本转语音增强系统的模型架构 tem

如图所示1，源音频和文本转语音音频被输入特征提取模块进行声学特征提取，随后由轻量级的基于双向长短期记忆网络的骨干网络进行面部动作预测。本研究采用混合变形格式来表示面部动作。最后，使用虚幻引擎4（UE4）将预测得到的混合变形序列渲染为动画 [1]。

增强音频-动画对的生成。增强后的音频-动画对通过以下方式生成：使用现成的文本转语音系统将源音频的脚本转换为文本转语音音频，并将与源音频匹配的混合形状作为文本转语音音频的标签。然而，文本转语音模型有其自身的时长预测器，导致源音频与文本转语音音频在词语级和序列级上的时长变化，从而难以对齐文本转语音音频与参考混合形状。因此，文本转语音音频中的每个声学特征帧所对应的参考混合形状并不像源音频-动画数据中的那样确定。文本转语音音频-动画对中的时长错位问题将在3.2中解决。

特征提取器。为更好地利用音频中蕴含的底层结构信息，我们采用预训练 HuBERT[16, 27] 模型作为特征提取器，并对各层的隐藏层输出进行加权求和，得到作为说话人脸系统输入的声学特征，可表示为：

$$f^t = \sum_{i=1}^{N}(\alpha_i * h_i^t) \quad , \quad \sum_{i=1}^{N}\alpha_i = 1$$

其中， h_i^t 表示 HuBERT 在时间 t 的第 i 层输出， N表示 HuBERT 中隐藏层的总数， α_i 表示 h_i^t的可学习权重， f^t 表示时间 t 的最终声学特征。

骨干网络。 我们首先采用一个卷积模块，将 HuBERT 提取的特征从 50Hz 下采样至 25Hz，以与混合形状对齐。该卷积下采样模块是一个由两层 1-D 卷积神经网络构成的模块。随后将下采样后的特征输入一个轻量级双向长短期记忆网络，该网络包含两层，之后通过一个线性层预测混合变形序列。

3.2. 用于文本转语音音频对齐的软动态时间规整损失

动态时间规整可以衡量两个不同长度序列之间的距离，因此适用于我们的任务，因为文本转语音音频与真实混合形状序列的长度不一定相同。然而，由于动态规划过程中包含取最小值操作，动态时间规整不可微。为得到可微的动态时间规整计算形式，我们采用了软动态时间规整 [14] 。软动态时间规整损失允许由音频生成的预测混合变形序列与真实混合形状序列进行软对齐，从而能够使用长度可变的文本转语音数据。据我们所知，这是该损失函数首次应用于说话人脸任务。

为构造软动态时间规整损失，我们首先考虑预测混合形状序列（$\hat{}b_1$, $\hat{}b_2$, $\cdots$, $\hat{}b_m$）与真实混合形状序列（b_1, b_2, $\cdots$, b_n）之间的动态时间规整距离。通过动态时间规整计算，两个序列中的词元将成对对齐；可行路径可以定义为满足某些限制条件的词元对有序序列，而动态时间规整的目标是

[1] 关于渲染的讨论不在本文讨论范围内，也不作详述 。

找到成本最低的可行路径。为获得动态时间规整的公式，设特定路径中的第 k 个配对项为 $\hat{}(b_{i_k},\ b_{j_k})$，动态时间规整受到以下两项限制：

$$\begin{aligned} i_{k+1} - i_k \leq 1, \quad j_{k+1} - j_k \leq 1 \quad \cdots\cdots continuity \\ i_{k+1} - i_k + j_{k+1} - j_k \geq 1 \quad \cdots\cdots monotonicity \end{aligned} \tag{1}$$

$\hat{}$将 δ_{ij} 记为词元 b_i 与 b_j 之间的距离。.在我们的工作中，采用欧几里得距离来度量两个词元之间的距离，因此 $\delta_{ij} = ||\hat{} b_i - b_j||_2$。可行路径的代价是该路径中所有配对词元之间距离的总和。为避免枚举所有可行路径，采用动态规划来寻找最优路径。将序列 $(\hat{}b_1, \cdots, \hat{}b_i)$ 与序列 $(b_1, \cdots, b_j)$ 之间最优动态时间规整路径的代价记为 $r_{i,j}$，我们的目标等价于求得最小的 $r_{m,n}$。根据公式1，最优路径代价的状态转移方程可表示为：

$$r_{i,j} = \delta_{ij} + \min\{r_{i-1,j}, r_{i,j-1}, r_{i-1,j-1}\} \tag{2}$$

最小值算子使 $r_{i,j}$ 不可微。为解决这一问题，软动态时间规整引入了最小值算子的平滑形式，即：

$$\min^{\gamma}\{d_1, \ldots, d_n\} := \begin{cases} \min_{i \leq n} d_i, & \gamma = 0 \\ -\gamma \log \sum_{i=1}^{n} e^{-d_i/\gamma}, & \gamma > 0 \end{cases} \tag{3}$$

当 $\gamma > 0$ 时，$\min^{\gamma}\{d_1, \ldots, d_n\}$ 可微；并且 γ 越接近 0，$\min^{\gamma}\{d_1, \ldots, d_n\}$ 就越接近 $\min\{d_1, \ldots, d_n\}$。公式 2 的软形式可由公式 3 推导，其中 $r_{i,j}$ 可微：

$$r_{i,\ j} = \delta_{ij} - \gamma \log(e^{-r_{i-1,\ j}/\gamma} + e^{-r_{i,\ j-1}/\gamma} + e^{-r_{i-1,\ j-1}/\gamma}) \tag{4}$$

最后，软动态时间规整损失的推导如下，$r_{m的计算n}$ 遵循式4中的状态转移方程，其时间复杂度为 $O(mn)$。

$$\mathcal{L}_{soft-DTW} = r_{m,n} \tag{5}$$

根据链式法则，$\frac{\partial \mathcal{L}}{\partial \delta_{ij}}$ 可由式4计算得出，更多关于梯度计算的数学细节请参阅 [14]。通过对 $\frac{\partial \mathcal{L}}{\partial \delta_{ij}}$ 进行反向传播，模型参数将在训练过程中更新。

4. 实验

4.1. 实验设置

表1. 数据集的说话人分配

配置名称	训练集	验证集	test
REC	录音	录音	录音
TTS1	1 个文本转语音说话人	1 个 TTS说话人	
TTS13	13 个 TTS说话人	13 个 TTS说话人	
REC+TTS13	录音+13TTS说话人	录音	

数据集描述 我们使用了一个来自世界人工智能创新（AIWIN）2022年文本与音频驱动说话人脸竞赛的中文数据集[2]。该数据集分为三个部分，包括训练集（276个序列，时长29分8秒）、验证集（79个序列，时长9分58秒）和测试集（85个序列，时长10分30秒）。该数据集中只有一位女性说话人（将该说话人的音频称为“录音”，其音频脚本可用），并使用 Microsoft Azure 的神经网络文本转语音系统生成文本转语音音频。实验采用了四种不同的数据集配置，分别为表格中的 REC、TTS1、TTS13 和 REC+TTS131，其中“录音”是与真实动画同步的录制音频，而文本转语音音频的时长可变，且与真实动画不同步。为进行公平比较，所有数据集配置均使用录制音频中的音频-动画数据作为测试集。

[2]http://ailab.aiwin.org.cn/competitions/69

实验设置 对于每项实验，我们将模型训练了 30 轮，并通过在验证集上评估软动态时间规整损失来选出最佳模型。最后，我们使用虚幻引擎4（UE4）将根据预测的混合形状生成的动画渲染为视频。演示可在我们的项目页面上查看。

评估指标 为分析实验结果，我们进行了客观评估和主观评估。在客观评估中，我们在“录制”测试集上测试模型的均方误差和动态时间规整分数。“录制”测试集中的音频和动画严格对齐，均方误差分数能够计算每对对齐词元之间距离的平均值，因此可在帧级量化序列的相似度。正如以往研究 [26]所证明的，动态时间规整能够发现时间序列中的模式，因此我们还引入了动态时间规整分数，以反映模型在序列级的整体表现。在主观评估中，我们通过 A/B 测试的形式开展用户研究。在 A/B 测试中，我们向志愿者展示了不同模型预测的动画，并请他们选出与背景音频同步效果更好的动画。A/B 测试所用的音频包括 2 段中文新闻报道音频、1 段中文“录制”音频、1 段中文文本转语音音频和 1 段英语音频。我们总共收集了 41 条反馈，以计算各模型的偏好率。

4.2. 文本转语音数据增强

在接下来的实验中，我们将探讨文本转语音数据增强对低资源场景的有效性。基线模型使用REC数据集中10个序列（时长70秒）进行训练。相比之下，我们根据10个序列的转录文本生成了另外13位说话人的TTS音频，并将其与录制音频结合，利用软动态时间规整算法增强基线说话人脸系统。实验结果见表2。

表2。使用录制音频、录制音频 + 13位说话人的TTS音频，在仅有70秒有标注数据的情况下训练所得模型的客观与主观结果

评估指标	REC	REC+TTS13
均方误差分数	0.00473	**0.00391**
动态时间规整分数	0.68819	**0.54433**
用户研究偏好率	31.0%	**69.0%**

结果表明，与基线系统相比，采用文本转语音数据增强的系统在客观和主观评估中均取得了稳定且显著的性能提升。我们的系统渲染效果也令人满意，这表明所提出的方法仅凭 70 秒的带标签音频-动画数据，就能构建出性能良好的音频驱动的说话人脸系统。更多信息请参阅项目页面上的演示。

4.3. 文本转语音驱动的说话人脸

接下来的实验旨在比较使用 TTS 音频训练的说话人脸系统与使用录制音频训练的系统的性能。为此，我们基于表 1 中定义的 REC、TTS1 和 TTS13 数据集训练了三个系统。实验结果如表3所示。

表3。在 29 分 8 秒训练数据上，使用录制音频、单个说话人的文本转语音音频以及 13 个说话人的文本转语音音频训练的说话人脸模型的客观和主观结果

评估指标	REC	TTS1	TTS13
均方误差分数	**0.00349**	0.00420	0.00378
动态时间规整分数	**0.46135**	0.53501	0.49961
用户研究偏好率	**61.0%**	-	39.0%

根据客观和主观结果，使用文本转语音数据训练的说话人脸模型略逊于使用录制音频训练的模型，这在预料之中，因为后者采用了同步音频，且模型是在匹配的录制音频上进行评估的。尽管存在这一差距，TTS13 的表现仍然令人瞩目，因为训练过程中完全没有使用同步录制语音。值得注意的是，TTS13 系统优于 TTS1 系统，这表明增加训练集的多样性有助于提升性能。文本转语音驱动的说话人脸系统能够取得令人满意的性能，这一结果从另一个角度证明了在音频驱动的说话人脸任务中结合 TTS 音频与软动态时间规整损失的有效性。

4.4. 模型鲁棒性

特征鲁棒性消融研究。所提出的方法采用 Hu-BERT 特征的加权和。为验证 HuBERT 特征的鲁棒性和优越性，我们比较了使用不同类型音频特征的说话人脸模型在不同有标注数据量下的表现。音频特征包括 HuBERT 特征、梅尔频率倒谱系数、音素后验概率图（PPGs）；其中，PPGs 是在中文语料库 AISHELL- 上训练的自动语音识别模型的最后一层输出2 [28]。结果如图 2所示。结果表明，我们的 HuBERT 特征显著且持续优于另外两种特征，取得了最佳表现。更多详情请参见项目页面。

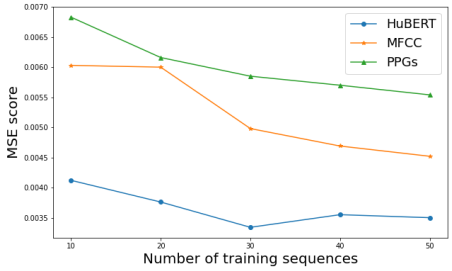


图 2。使用 HuBET、梅尔频率倒谱系数、音素后验概率图特征训练的模型的均方误差分数，数据量为 10、20、30、40、50 个训练序列

损失函数有效性消融研究。接下来的实验旨在证明，与说话人脸任务中常用的传统 L1 和 L2 损失相比，软动态时间规整损失的有效性。实验期间，数据集配置固定为 REC。采用不同损失函数训练的模型的客观评估结果如表4所示。我们发现，采用 L1、L2 和软动态时间规整训练的模型，其均方误差分数相近，这表明在帧级上，软动态时间规整损失的拟合性能与 L1 和 L2 损失相当。同时，采用软动态时间规整训练的模型在动态时间规整分数上远胜于采用 L1 和 L2 损失训练的模型，这表明它有潜力更好地预测形状与真实值相似的序列。为了比较拟合序列形状的能力，我们从测试集中随机选取一段音频，并绘制“下颌张开”值（混合形状的一个维度）。图3显示，采用软动态时间规整训练的模型更擅长预测混合变形序列中的峰值，因此有望生成表现力更丰富的渲染动画。

表 4。使用 L1、L2 和软动态时间规整损失训练的模型性能

评估指标	L1	L2	软动态时间规整
均方误差分数	**0.00344**	0.00351	0.00349
动态时间规整分数	0.51082	0.55999	**0.46135**

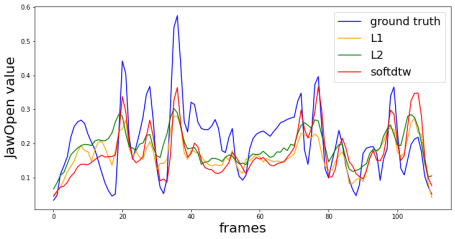


图 3。真实值与使用 L1、L2 和软动态时间规整损失训练的模型预测值的下颌张开变化趋势

模型架构消融研究。为了进一步验证我们方法的有效性，我们还将骨干网络从双向长短期记忆网络改为门控循环单元、循环神经网络和 Transformer，并分别训练了使用或不使用文本转语音数据增强的模型。结果如表 5所示，文本转语音数据增强方法在全部 4 种架构上均表现出优势。

表 5。使用录制音频、录制音频 + 和 13 位说话者的文本转语音音频训练的不同网络架构模型的均方误差分数，仅使用 70 秒有标注数据

	双向长短期记忆网络	GRU	RNN	Transformer
不使用数据增强	0.00473	0.00438	0.00447	0.00896
使用数据增强	**0.00391**	**0.00406**	**0.00409**	**0.00632**

5. 结论

我们提出了一种用于说话人脸任务的文本转语音数据增强方法，通过文本转语音系统生成增强的音频-动画对，并引入软动态时间规整损失，解决了文本转语音音频带来的错位问题。我们采用 Hu-BERT 特征的加权和，以充分利用音频的底层信息。客观实验和主观实验结果证明，所提出的方法能够在数据资源有限的情况下提升说话人脸系统的小样本能力。未来，我们计划将该文本转语音数据增强方法应用于说话人脸生成的其他场景，例如合成照片级真实感说话人脸。我们希望该方法能够成为提升说话人多样性和说话人脸任务小样本能力的一种替代方案。

致谢： 本研究得到中国国家社会科学基金（编号：6220070337 和 18ZDA293）、媒体融合生产技术与系统国家重点实验室项目（编号：SKLMCPTS 2020003）、上海市科技重大项目（2021 SHZDZX0102）及鹏城实验室国际合作项目的资助。

6. 参考文献

[1] Justus Thies、Michael Zollhofer、Marc Stamminger、Christian Theobalt 和 Matthias Nießner，《Face2face：RGB 视频的实时人脸捕捉与表演重演》，载于*IEEE* 计算机视觉与模式识别会议论文集，2016年。[2] Michael Zollh¨ofer、Justus Thies、Pablo Garrido、Derek Bradley、Thabo Beeler、Patrick P´erez、Marc Stamminger、Matthias Nießner 和 Christian Theobalt，《单目 3D 人脸重建、跟踪与应用的最新进展》，载于计算机图形论坛。Wiley Online Library，2018年。[3] Koki Nagano、Jaewoo Seo、Jun Xing、Lingyu Wei、Zimo Li、Shunsuke Saito、Aviral Agarwal、Jens Fursund、Hao Li、Richard Roberts 等，《PAGAN：使用动态纹理的实时虚拟形象》，*ACM* 图形学汇刊，2018年。[4] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner，《神经语音木偶操控：音频驱动的人脸表演重演》，载于欧洲计算机视觉会议。Springer，2020年。[5] Liyang Chen、Zhiyong Wu、Jun Ling、Runnan Li、Xu Tan 和 Sheng Zhao，《Transformer-S2A：稳健高效的语音转动画》，载于*ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年。[6]Yang Zhou、Xintong Han、Eli Shechtman、Jose Echevarria、Evangelos Kalogerakis 和 Dingzeyu Li，《MakeItTalk：说话人感知的说话头动画》，*ACM* 图形学汇刊（*TOG*），2020年。[7] Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J. Black，《3D 说话风格的捕捉、学习与合成》，载于*2019 IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），2019年。[8] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang 和 Taku Komura，《Faceformer：基于 Transformer 的语音驱动 3D 面部动画》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年。[9] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，*arXiv* 预印本 *arXiv:2006.04558*，2020年。[10] Chenpeng Du 和 Kai Yu，《基于音素级混合密度网络的丰富韵律多样性建模》，*arXiv* 预印本 *arXiv:2102.00851*，2021年。[11] Chenpeng Du、Yiwei Guo、Xie Chen 和 Kai Yu，《VQTTS：基于自监督 VQ 声学特征的高保真文本转语音合成》，*arXiv* 预印本 *arXiv:2204.00768*，2022年。[12] Chenpeng Du 和 Kai Yu，《面向低资源语音识别的说话人增强》，载于*ICASSP 2020-2020 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年。[13] Chenpeng Du、Bing Han、Shuai Wang、Yanmin Qian 和 Kai Yu，《SynAug：面向文本相关说话人验证的基于合成的数据增强》，载于*ICASSP 2021-2021 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年。[14] Marco Cuturi 和 Mathieu Blondel，《Soft-DTW：一种用于时间序列的可微损失函数》，载于国际机器学习会议。PMLR，2017年。

[15] Hiroaki Sakoe，《用于连续语音识别的动态规划方法》，载于1971年《国际声学大会论文集》，布达佩斯，1971年。[16]Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov和Abdelrahman Mohamed，《HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习》，《IEEE/ACM音频、语音与语言处理汇刊》，2021年。[17] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri和CV Jawahar，《只需一个唇同步专家，即可实现真实场景下的语音到唇部动作生成》，载于《第28届ACM国际多媒体会议论文集》，2020年。[18] Awni Hannun、Carl Case、Jared Casper、Bryan Catanzaro、Greg Diamos、Erich Elsen、Ryan Prenger、Sanjeev Satheesh、Shubho Sengupta、Adam Coates等，《Deep Speech：扩展端到端语音识别》，arXiv预印本 arXiv:1412 5567，2014年。[19] Kaizhi Qian、Yang Zhang、Shiyu Chang、Xuesong Yang和Mark Hasegawa-Johnson，《AutoVC：仅凭自编码器损失实现零样本语音风格迁移》，载于《国际机器学习会议》。PMLR，2019年。[20]Steffen Schneider、Alexei Baevski、Ronan Collobert和Michael Auli，《wav2vec：用于语音识别的无监督预训练》，《Interspeech论文集》，2019年。[21] Alexei Baevski、Steffen Schneider和Michael Auli，《vq-wav2vec：离散语音表征的自监督学习》，arXiv预印本 arXiv:1910.05453，2019年。[22] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed和Michael Auli，《wav2vec 2.0：自监督语音表征学习框架》，《神经信息处理系统进展》，2020年。[23] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatong Gu和Michael Auli，《data2vec：用于语音、视觉和语言自监督学习的通用框架》，《国际机器学习会议》，2022年。[24] Ziyang Ma、Zhisheng Zheng、Changli Tang、Yujin Wang和Xie Chen，《MT4SSL：整合多目标以提升自监督语音表征学习》，arXiv预印本 arXiv:2211.07321，2022年。[25] Shu-wen Yang、Po-Han Chi、Yung-Sung Chuang、Cheng-I Jeff Lai、Kushal Lakhotia、Yist Y Lin、Andy T Liu、Jiatong Shi、Xuankai Chang、Guan-Ting Lin等，《SUPERB：语音处理通用性能基准》，arXiv预印本 arXiv:2105.01051，2021年。[26] Donald J Berndt和James Clifford，《利用动态时间规整在时间序列中寻找模式》，载于KDD研讨会，1994年。[27]《腾讯游戏语音中文语音预训练模型》，https://github.com/TencentGameMate/chinese speech pretrain，2022年。[28] Jiayu Du、Xingyu Na、Xuechen Liu和Hui Bu，《AISHELL-2：推动普通话自动语音识别研究迈向工业级规模》，arXiv预印本 arXiv:1808.10583，2018年。